ECOGRAPHY 25: 102–108, 2002

On the estimation of species richness based on the accumulation of previously unrecorded species

Emmanuelle Cam, James D. Nichols, John R. Sauer and James E. Hines

Cam, E., Nichols, J. D., Sauer, J. R. and Hines, J. E. 2002. On the estimation of species richness based on the accumulation of previously unrecorded species. – Ecography 25: 102–108.

Estimation of species richness of local communities has become an important topic in community ecology and monitoring. Investigators can seldom enumerate all the species present in the area of interest during sampling sessions. If the location of interest is sampled repeatedly within a short time period, the number of new species recorded is typically largest in the initial sample and decreases as sampling proceeds, but new species may be detected if sampling sessions are added. The question is how to estimate the total number of species. The data collected by sampling the area of interest repeatedly can be used to build species accumulation curves: the cumulative number of species recorded as a function of the number of sampling sessions (which we refer to as ''species accumulation data''). A classic approach used to compute total species richness is to fit curves to the data on species accumulation with sampling effort. This approach does not rest on direct estimation of the probability of detecting species during sampling sessions and has no underlying basis regarding the sampling process that gave rise to the data. Here we recommend a probabilistic, nonparametric estimator for species richness for use with species accumulation data. We use estimators of population size that were developed for capture-recapture data, but that can be used to estimate the size of species assemblages using species accumulation data. Models of detection probability account for the underlying sampling process. They permit variation in detection probability among species. We illustrate this approach using data from the North American Breeding Bird Survey (BBS). We describe other situations where species accumulation data are collected under different designs (e.g., over longer periods of time, or over spatial replicates) and that lend themselves to of use capture-recapture models for estimating the size of the community of interest. We discuss the assumptions and interpretations corresponding to each situation.

E. Cam (ecam@sfu.ca), Dept of Forestry, North Carolina State Univ., U.S. Geological Survey, Biological Resources Div., Patuxent Wildlife Research Center, 11510 American Holly Dr., Laurel, MD 20708-4019, USA (present address: Dept of Biological Sciences, Simon Fraser Univ., 8888 University Dr., Burnaby, B.C., Canada V5A 1S6). – J. D. Nichols, J. R. Sauer and J. E. Hines, U.S. Geological Survey, Biological Resources Div., Patuxent Wildlife Research Center, 11510 American Holly Dr., Laurel, MD 20708-4019, USA.

Estimation of species richness of local communities or of groups defined taxonomically or ecologically is an important step for investigations in community ecology and conservation biology. Investigation of variation in species richness over space and time, and testing hypotheses about factors potentially associated with these variations, are major research directions in those fields. As emphasized by Colwell and Coddington (1994), the importance and urgency of the task of evaluating biodiversity require that we devote substantial effort to estimation of species richness. Estimation of total richness of a community of interest is not trivial. Indeed, situations where investigators can build exhaustive lists of species present in the area of interest are rare,

Accepted 5 June 2001

ISSN 0906-7590

especially when focusing on animals. Observed richness based on species counts over limited time periods often underestimates actual richness. This problem has motivated substantive efforts to develop approaches to estimation of species richness in the face of species detection probabilities that are variable and < 1.

Here we consider a particular sampling design and estimate the size of species assemblages using estimators originally developed to estimate population size from capture-recapture data. We focus on situations where investigators sample the same area repeatedly within a relatively short period of time. Either investigators record all the species that they detect in each sampling effort, or they exclusively record "new species" ("new" relative to those already recorded during previous sampling sessions). In both cases, it is possible to plot the cumulative number of species detected as a function of the number of units of effort expended, such as the number of sampling sessions, or time. The resulting curve is a species accumulation curve or "collector's curve" (Soberón and Llorente 1993, Colwell and Coddington 1994). The only information needed to build such curves is the first sampling occasion where the species is recorded, that is data from "new" species exclusively. For the general sampling situation in which the investigator records all species detected at each sampling occasion, we recommend consideration of the full set of models available for estimation of species richness (e.g., Otis et al. 1978, Bunge and Fitzpatrick 1993, Colwell and Coddington 1994, Lee and Chao 1994, Nichols and Conroy 1996, Boulinier et al. 1998, Chazdon et al. 1998). In this note, we focus on methods for use in situations where the only data available are the number of new species recorded in each sampling session.

The number of new species recorded in each unit of sampling effort tends to decrease as the number of units of effort increases. That is, species accumulation curves usually flatten when sampling effort is sufficiently large. However, even in the "flat" portion of curves, new species may be detected if sampling sessions are added. The question is how to estimate the total number of species from accumulation data, given that direct observation seldom permits access to the complete list of species for the site considered. Soberón and Llorente (1993) and Colwell and Coddington (1994) reviewed several parametric methods to extrapolate species accumulation data. Colwell and Coddington (1994) also reviewed nonparametric methods for estimating species richness from samples, but these reviewed nonparametric approaches all use methods based on the full detection history data (presence/absence of a species in each sampling unit; Nichols and Conroy 1996), rather than on species accumulation data. In this note, we recommend a probabilistic nonparameteric estimator for species richness for use with species accumulation data. We focus on direct estimation approaches accounting for the underlying sampling process.

Estimation models for the accumulation of previously unrecorded species

Most of the nonparametric methods for estimating species richness from full detection history data arise from the modeling of capture-recapture data for animal populations (Otis et al. 1978, Bunge and Fitzpatrick 1993, Colwell and Coddington 1994, Nichols and Conroy 1996). The class of capture-recapture models of particular relevance for use with species accumulation data are the "behavioral response" and "removal" models. Behavioral response models consider the situation where an animal has one capture probability before it has ever been captured and another capture probability after it has been caught for the first time. The sufficient statistics for estimation of population size under these models are the number of new captures at each sampling period (Otis et al. 1978). Removal models are needed in situations where animals are removed from the population at initial capture (e.g., some sampling methods such as snap-trapping for small mammals and electrofishing for aquatic vertebrates kill animals as part of the capture process), so the only statistics available for estimation are again the numbers of captures (all are of "new" animals) at each sample period. This sampling situation is thus analogous to community-level sampling in which only new species are recorded at each sampling occasion.

Two basic models are available for use with removal data, and thus with species accumulation data, models M(b) and M(bh), where "b" denotes "behavior", and "h" stands for "heterogeneity" (Otis et al. 1978, Pollock et al. 1990, Nichols and Conroy 1996, Boulinier et al. 1998). Note that the label "b" does not restrict use of these models to animals (these can be used for plants species as well), but instead indicates a distinction between initial and subsequent detection and an estimation focus on initial detections (i.e., on species accumulation data). Model M(b) assumes that all species in the community have identical detection probabilities, whereas model M(bh) permits each species to have a different detection probability. Models permitting heterogeneous detection probabilities among species have been found to be especially useful in previous modeling of community data (Boulinier et al. 1998). Such heterogeneity can be caused by behavioral differences among species as well as by differences in the abundance of individuals within the different species (Alpizar-Jara et al. unpubl.). Despite the likely appropriateness of model M(bh) for many data sets, we still believe that it is reasonable to include consideration of model M(b). Although species detection probabilities

will likely never be identical for any group of species, it is possible that they could be sufficiently similar to be most parsimoniously modeled using a single parameter (e.g., see discussion of the principle of parsimony in Burnham and Anderson 1998).

Maximum likelihood estimates under model M(b) are provided by program CAPTURE (Otis et al. 1978, Rexstad and Burnham 1991). Standard maximum likelihood estimation is not possible under the general M(bh) model, as there are too many parameters to estimate (a detection probability for each species). Several reasonable estimators have been developed for model M(bh), however, including the generalized removal estimator of Otis et al. (1978), the jackknife estimator of Pollock and Otto (1983), the coverage estimator of Lee and Chao (1994), and the nonparametric maximum likelihood approach Norris and Pollock (1996). The generalized removal estimator and the Pollock and Otto (1983) jackknife are both computed by program CAPTURE. Although both of these estimators have performed well in simulation studies (Pollock and Otto 1983, Lee and Chao 1994), the Pollock and Otto (1983) jackknife estimator performed better in the case of high heterogeneity, so may be the best choice for species accumulation data.

The jackknife estimator of Pollock and Otto (1983) is simply:

$$\hat{N} = \sum_{i=1}^{K-1} u_i + K u_K,$$

where N is species richness, K is the number of sampling occasions, and u_i is the number of new (previously undetected) species detected at sampling occasion i. The form of the variance estimator for $\hat{N}$ is provided by Pollock and Otto (1983). The 95% confidence interval for the richness estimate is computed by assuming a lognormal distribution for the estimated number of species not detected, as recommended by Chao (1989) and Rexstad and Burnham (1991).

In addition to computing these estimates, CAPTURE includes a model selection algorithm based on a discriminant function developed using simulated data under all 8 basic models (Otis et al. 1978). If just species accumulation data are available, then this selection algorithm can be used to decide whether model M(b) or model M(bh) is most likely to be useful.

Example analyses

Here, we illustrate use of these models for estimating species richness from species accumulation data. We use avian point count data, more precisely data collected during research on the North American Breeding Bird Survey (BBS; Robbins et al. 1986, Sauer et al. 1997). The BBS is a roadside survey conducted in the spring. Experienced BBS observers were asked to conduct multiple surveys of selected sites in the spring of 1991 (Link et al. 1994). The data were collected in Alabama, Louisiana, Maine, Maryland, New Hampshire, New Mexico and Vermont. We used data on only the initial detections of species from 9 of these routes in order to place ourselves in the framework of analysis of species accumulation data.

We used program CAPTURE (Rexstad and Burnham 1991) to compute statistics useful in model selection and estimates of species richness. Because of our restriction to new detections (accumulation data), we focused only on the relative discriminant function scores for models M(b) and M(bh). We used the jackknife estimator (Pollock and Otto 1983) for model M(bh) as this estimator has performed well in simulation studies, as mentioned above.

Model M(bh) received a higher discriminant function score than model M(b) in all 9 data sets, as expected. We thus restrict interest to model M(bh) estimates (Table 1). The jackknife estimator of species richness proposed by Pollock and Otto (1983) for model M(bh) leads to estimates different from observed total richness, S, only when at least one new species is recorded

Table 1. Observed and estimated species richness.

Route	Number of sessions	S	$\hat{N}$	$\hat{SE}$ ($\hat{N}$)	CI ($\hat{N}$)
58017	11	87	97	10.48	89–141
46500	13	86	97	12.49	88–149
42105	12	54	65	11.49	57–113
60027	7	33	33	0	
2043	11	75	95	14.83	81–148
2017	11	75	75	0	
44041	5	87	99	7.75	91–125
87005	11	78	78	0	
46039	13	80	80	0	

S: observed total species richness
$\hat{N}$: estimated species richness
$\hat{SE}$: estimated standard error
CI: 95% confidence interval

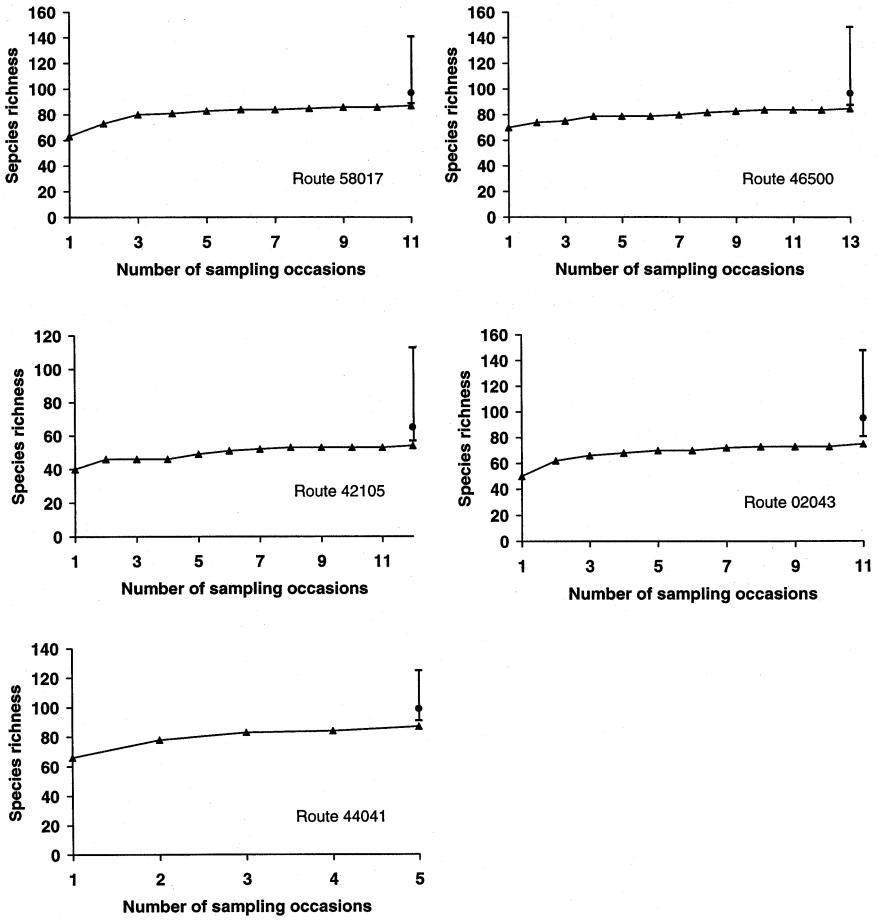

Fig. 1. Cumulative observed species richness (▲) as a function of the number of sampling sessions. Estimate of species richness (●) and 95% confidence interval (model M(bh)).

in the last occasion. In 5 of 9 routes in this study, estimates of species richness are higher than observed total richness (Fig. 1).

Discussion

Our example data come from the same observers sampling the same locations at multiple occasions occurring within a short time period (all sampling within 3–4 weeks) during which the community composition was expected to remain the same. Observers expended the same effort per occasion on all sampling occasions. Data of this sort are ideally suited for use with the capture-recapture approaches of models M(b) and M(bh).

Models M(b) and M(bh) may also be useful for other sampling approaches. For example, species accumulation data are sometimes collected over spatial, rather than temporal, replicates (e.g., Flather 1996). Assume a situation where we are interested in a community of species associated with some large area of interest. Further assume that we select sampling plots randomly from this large area, and that we then record the number of new species detected at each new sample location, visiting all locations within a relatively short period of time (so that the community does not change) and attempting to expend the same effort at each location. Species accumulation data from such a sampling scheme can also be used with model M(bh). Now the species detection probability associated with the modeling represents the product of two probabilities:

$$\Pr(\text{at least 1 individual of species is present in sample area} \mid \text{presence in large area}) \times \Pr(\text{detection} \mid \text{at least 1 individual of species is present in sample area}).$$

The resulting estimate of species richness should correspond in this situation to the total number of species in the large area of interest.

Still another sampling situation for which model M(bh) might be useful involves sampling over long time periods such as several years (e.g., Soberón and Llorente 1993). In this case, an area is sampled each year, perhaps, and species accumulation data collected. Although the conceptual framework underlying such studies is not often specified, we deduce that the "community" of interest in such studies must be viewed as a group of species with some nonnegligible probability of being present in an area in any given year or sampling period. We might refer to such a community as a "supercommunity" (analogous to the superpopulation concept of Crosbie and Manly 1985, Kendall et al. 1997). We can then view the actual species present in the area in a particular year as a stochastic selection from this supercommunity. The mechanistic process that produces each year's selection of species involves local species extinctions and colonizations, but the details of the stochastic process need not be specified. Under the assumptions that the different species can have different probabilities of being present and then detected, but that each individual species has the same underlying probability of being present and being detected during any year of the study, then model M(bh) should be an appropriate model. The detection probability under this sampling situation becomes the product of 2 probabilities:

$$\Pr(\text{species present during sampling period of year i} \mid \text{member of supercommunity}) \times \Pr(\text{detection} \mid \text{species present during sampling period of year i}).$$

The assumption likely to limit utility of this approach concerns the absence of temporal variation in the probability of presence within a species. In particular, we might expect that the probability that a species is present during a particular year in an area of interest might vary as a first-order Markov process (see related discussion in Kendall et al. 1997). That is, we might expect probability of presence in year i to be different depending on whether or not the species was present in year $i-1$. Thus, we conclude that model M(bh) may be useful in estimating species richness for areas based on multiple years of sampling, but that the assumptions required in this case are restrictive and merit careful examination.

Finally, consider sampling conducted at a single location over a short time period (e.g., a few weeks) such as our BBS examples, but with unequal effort expended at the different sampling occasions. If the sampling effort can be quantified (e.g., as number of person-days, trap-nights, net-hours, etc., Soberón and Llorente 1993), then we can use a catch-effort removal model (e.g., Seber 1982, Pollock et al. 1984, Lee and Chao 1994, Gould and Pollock 1997) for estimation of species richness from data on number of new species detected and number of units of effort expended for each sample occasion.

The question of "stopping rules" specifying when to stop sampling arises naturally in the consideration of species accumulation data. We suspect that many studies will involve a fixed number of sampling periods, so that the question of stopping rules will not occur. Nevertheless, other sampling programs may involve repeated sampling with no fixed endpoint, and stopping rules would be useful in these situations. Stopping rules must be based on a specified objective function for the sampling effort. For example, in the subject area of software testing, the objective function weighs the cost of undetected "bugs" against the cost of further sampling and testing (see Chao et al. 1993). If species sampling is embedded in a program with specific scientific or management objectives, then these objectives can be used to develop a suitable objective function from which stopping rules can be derived. In the absence of such specific objectives, stopping rules for species accumulation studies might be based on a measure of precision for the estimate of species richness (e.g., coefficient of variation):

$$\hat{CV}(\hat{N}_k) = \frac{\hat{SE}(\hat{N}_k)}{\hat{N}_k}$$

where k denotes the number of the occasion of sampling for which the stopping rule is to be assessed, or on the proportion of species detected, P_k:

$$\hat{P}_k = \frac{S_k}{\hat{N}_k}$$

where $S_k = \Sigma_{i=1}^{k} u_i =$ total species detection through sample occasion k (e.g., we might stop sampling when the proportion of species detected exceeds some arbitrary threshold such as 90%). The Pollock and Otto (1983) estimator yields $\hat{N}_k = S_k$ (thus $\hat{P}_k = 1$) for $u_k = 0$, but we do not believe that this should necessarily lead an investigator to stop sampling on the first occasion at which no new species are encountered. Because sampling is viewed as a probabilistic process, it is probably not wise to attach too much significance to single $u_i = 0$.

There is an extensive literature on the various parametric functions used to fit curves describing species accumulation with sampling effort, whether effort corresponds to the number of units of space, of time, or the number of samples collected (e.g., Connor and McCoy 1979, McGuinness 1984, Flather 1996). Although in most cases investigators have recorded the total number of species observed at each sampling occasion or in each unit of area sampled (i.e., they have not restricted data collection to previously unrecorded species only), the analogy with the situation addressed here is clear: extrapolation of curves based on observed

species richness using parametric methods is a common approach that has a long history. Although these approaches were developed in response to the recognition that observed species richness can be a biased estimate of actual richness because of the failure to detect all species in sampling efforts, most of the functions used in these approaches were not developed from underlying models of detection probability. Instead, functions appear to have been selected because of their ability to assume shapes characteristic of observed species accumulation data. Unfortunately, it is common for several competing parametric functions to fit a particular set of species accumulation data well (providing little basis for selecting among them), yet yield very different estimates of richness (i.e., have different asymptotes; see Soberón and Llorente 1993, Colwell and Coddington 1994, Flather 1996). Although recent information-theoretic approaches to model selection (Burnham and Anderson 1998) should at least provide an objective basis for selecting a function describing accumulation data, these approaches are conditional on a reasonable a priori model set. We prefer that a model set developed for use with species accumulation data includes mechanistic and probabilistic models of detection probability, such as models M(bh) and M(b).

In summary, because model M(bh) was developed specifically for the sampling process underlying data of the sort represented by species accumulation curves, we believe that it should be the logical "first choice" for estimating species richness using species accumulation data from samples of equal effort over short periods of time. In particular, this direct estimation approach should be preferable to attempts to estimate the asymptote of phenomenological models with no underlying mechanistic basis regarding the sampling process. If the data are accumulated over a long period of time (e.g., many years), then model M(bh) may be useful in situations for which the assumption of equal probability of presence and detection over time is reasonably met. In sampling situations with unequal, but known, sampling effort devoted to the different sampling occasions, closed-population catch-effort models (Seber 1982, Pollock et al. 1984, Gould and Pollock 1997) should be useful. Finally, we remind readers that these recommendations are for species accumulation data only. If the full detection history data (species lists for all sampling periods, rather than just new species detected) are available, then we recommend the use of the full set of models implemented in program CAPTURE (Nichols and Conroy 1996, Boulinier et al. 1998).

Acknowledgements – E. Cam was supported by a Cooperative Agreement between the United States Dept of Agriculture's Forest Service and North Carolina State Univ. Dept of Forestry. Program CAPTURE is available at ⟨http://www.mbr-pwrc.usgs.gov⟩. We thank R. Alpizar-Jara and K. H. Pollock for helpful discussions, and R. K. Colwell for constructive comments on earlier versions of this paper.

References

Boulinier, T. et al. 1998. Estimating species richness to make inferences in community ecology: the importance of heterogeneity in species detectability as shown from capture-recapture analyses of North American Breeding Bird Survey data. – Ecology 79: 1018–1028.

Bunge, J. and Fitzpatrick, M. 1993. Estimating the number of species: a review. – J. Am. Stat. Assoc. 88: 364–373.

Burnham, K. P. and Anderson, D. R. 1998. Model selection and inference, a practical information-theoretic approach. – Springer.

Chao, A. 1989. Estimating population size for sparse data in capture-recapture experiments. – Biometrics 45: 427–438.

Chao, A., Ma, M.-C. and Yang, M. C. K. 1993. Stopping rules and estimation for recapture debugging with unequal failure rates. – Biometrika 80: 193–201.

Chazdon, R. L. et al. 1998. Statistical methods for estimating species richness of woody regeneration in primary and secondary rain forests of the NE Costa Rica. – In: Dallmeir, F. and Comiskey, J. A. (eds), Forest biodiversity research, monitoring and modeling: conceptual background and Old World case studies. Parthenon Publ., Paris, pp. 285–309.

Colwell, R. K. and Coddington, J. A. 1994. Estimating terrestrial biodiversity through extrapolation. – Philos. Trans. R. Soc. Lond. B 345: 101–118.

Connor, E. F. and McCoy, E. D. 1979. The statistics and biology of the species-area relationship. – Am. Nat. 113: 791–833.

Crosbie, S. F. and Manly, B. F. J. 1985. Parsimonious modelling of capture-mark-recapture studies. – Biometrics 41: 385–398.

Flather, C. H. 1996. Fitting species-accumulation functions and assessing regional land use impacts on avian diversity. – J. Biogeogr. 23: 155–168.

Gould, W. R. and Pollock, K. H. 1997. Catch-effort maximum likelihood estimation of important population parameters. – Can. J. Fish. Aquat. Sci. 54: 890–897.

Kendall, W. L., Nichols, J. D. and Hines, J. E. 1997. Estimating temporary emigration using capture-recapture data with Pollock's robust design. – Ecology 78: 563–578.

Lee, S.-M. and Chao, A. 1994. Estimating population size via sample coverage for closed capture-recapture models. – Biometrics 50: 88–97.

Link, W. A. et al. 1994. Within-site variability in surveys of wildlife populations. – Ecology 75: 1097–1108.

McGuinness, K. A. 1984. Equations and explanations in the study of species-area curves. – Biol. Rev. 59: 423–440.

Nichols, J. D. and Conroy, M. J. 1996. Estimation of species richness. – In: Wilson, D. E. et al. (eds), Measuring and monitoring biological diversity. Standard methods for mammals. Smithsonian Inst. Press, USA, pp. 226–234.

Norris, J. L. III and Pollock, K. H. 1996. Nonparametric MLE under two closed capture-recapture models with heterogeneity. – Biometrics 52: 639–649.

Otis, D. L. et al. 1978. Statistical inference from capture data on closed animal populations. – Wildl. Monogr. 62.

Pollock, K. H. and Otto, M. C. 1983. Robust estimation of population size in closed animal populations from capture-recapture experiments. – Biometrics 39: 1035–1050.

Pollock, K. H., Hines, J. E. and Nichols, J. D. 1984. The use of auxiliary variables in capture-recapture and removal experiments. – Biometrics 40: 329–340.

Pollock, K. H. et al. 1990. Statistical inference for capture-recapture experiments. – Wildl. Monogr. 107.

Rexstad, E. and Burnham, K. P. 1991. User's guide for interactive program CAPTURE. Abundance estimation of closed animal populations. – Colorado State Univ.
Robbins, C. S., Bystrak, D. and Geissler, P. H. 1986. The breeding bird survey: its first fifteen years, 1965–1979. – U.S. Fish Wild. Serv. Resour. Publ. 157.
Sauer, J. R. et al. 1997. The North American breeding bird survey results and analysis. Ver. 96.4. – Patuxent Wildlife Research Center, Laurel, MD.
Seber, G. A. F. 1982. Estimation of animal abundance and related parameters. – Macmillan.
Soberón, J. M. and Llorente, J. B. 1993. The use of species accumulation functions for the prediction of species richness. – Conserv. Biol. 7: 480–488.